Concentration inequalities, counting processes and adaptive statistics

Patricia Reynaud-Bouret

ESAIM: PROCEEDINGS, Vol. ?, 2013, 1-10
Editors: Will be set by the publisher

CONCENTRATION INEQUALITIES, COUNTING PROCESSES AND ADAPTIVE STATISTICS *

Patricia Reynaud-Bouret [1]

Abstract. Adaptive statistics for counting processes need particular concentration inequalities to define and calibrate the methods as well as to precise the theoretical performance of the statistical inference. The present article is a small (non exhaustive) review of existing concentration inequalities that are useful in this context.

Résumé. Les statistiques adaptatives pour les processus de comptage nécessitent des inégalités de concentration particulières pour définir et calibrer les méthodes ainsi que pour comprendre les performances de l'inférence statistique. Cet article est une revue non exhaustive des inégalités de concentration qui sont utiles dans ce contexte.

Mathematics Subject Classification: 62G05, 62G10, 62M07, 62M09, 60G55, 60E15.

Keywords: Point process; counting process; adaptive estimation; adaptive testing; Talagrand type inequalities; Bernstein type inequalities.

Introduction

Counting processes have been used for many years to model a large variety of situations from biomedical data to seismic or financial data [2,34,35]. In neuroscience, they have been used almost as soon as spike trains have been recorded [38]. They more recently appeared in genomics, for instance as a limit of word distribution on the DNA [40]. Depending on the kind of data, it is difficult to have a priori knowledge on their distribution and it is possible that those distributions are very irregular, as we will see later. Therefore there is a huge need for adaptive statistical methods that are able to deal with those kind of distributions.

The recent advances of adaptive statistics have been due to a strong link between the calibration of those methods and concentration inequalities (see for instance [32]). It is the aim of the present article to list some concentration inequalities for counting processes that are useful when dealing with adaptive statistics.

Before going any further, let us define our main objects. A *point process*, N, is a random countable set of points of some measurable space $\mathbb{X}$. We denote by N_A, the number of points of N in the set A and dN denotes the point measure i.e. the sum of the Dirac masses at each point of N (see [15] for a complete overview).

Poisson processes are the simplest point processes one can encounter [28]:

Definition 1. *A Poisson process, N, on a measurable space $\mathbb{X}$ is a random countable set of points such that*

* *This research is partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration) and by the PEPS BMI 2012-2013 Estimation of dependence graphs for thalamo-cortical neurons and multivariate Hawkes processes.*

[1] Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, 06100 Nice, France.

- *for all integer n, for all $A_1, \dots, A_n$ disjoint measurable subsets of $\mathbb{X}$, $N_{A_1}, \dots, N_{A_n}$ are independent random variables.*
- *for all measurable subset A of $\mathbb{X}$, N_A obeys a Poisson distribution with parameter depending on A and denoted $\mathfrak{m}(A)$.*

In the above definition, $\mathfrak{m}$ actually defines a measure on $\mathbb{X}$, which is called the mean measure of the process N. Usually, $\mathbb{X}$ is a subset of $\mathbb{R}^d$ and $\mathfrak{m}$ is absolutely continuous with respect to the Lebesgue measure. In this case, $s(x) = d\mathfrak{m}/dx$ is the intensity of the Poisson process N.

Poisson processes model a large variety of situations, and let us cite, as examples, positions of genes on the DNA sequence [40], size of individual oil reservoirs in some petroleum field [54], stockprice changes of extraordinary magnitude [34]. All these data have something in common: there are good evidences that the underlying intensity is highly irregular, with very localized spikes with unknown positions and shapes and, for some of them, a very large and typically unknown support, when this support is finite.

The adaptive statistical inference aims at estimating s or at testing some hypothesis on s with as few assumptions on s (or on the alternatives) as possible. Typically s belongs to $\mathbb{L}_p(\mathbb{X})$, for $1 \leq p \leq +\infty$, but we do not want to assume that s is smooth with the precise known regularity α. Since the regularity is unknown and since the rate of convergence (or rate of separation) depends on the regularity, we want the procedures to adapt to this unknown regularity and to be as precise as the methods knowing the precise regularity of the target function s.

Actually, exhibiting such kind of procedures is not just a pure theoretical game. The procedures that are theoretically adaptive, if they are calibrated, may provide practical methods that are performing really well and that are robust to changes in the target function s. Indeed an adaptive method that does not need any information on the regularity or on the support of the target function is easier to use since we do not need to ask the practitioner to provide such information on s before proceeding to the inference.

The calibration step, as we will see later, is at the root of the need for concentration inequalities in adaptive statistics.

Counting processes are a generalization of the notion of Poisson process on the real (positive) line. If the point process is, say, almost surely finite in finite intervals (i.e. $N_{[a,b]} < \infty$ a.s.), then one can count the points: if the positive real line represents time after 0, then there exist a first point, a second .... The random function $N_t = N_{[0,t]}$ as a function of $t \in \mathbb{R}$ is a piecewise constant increasing function such that $N_0 = 0$ with jumps equal to 1. This is the definition of a *counting process*. The interested reader may find in [9] a complete review of those processes and precise definitions. Under suitable assumptions, one can informally define the (conditional) intensity $\lambda(t)$ of the counting process $(N_t)_{t\geq 0}$ by

$$\mathbb{E}(dN_t|\mathcal{F}_{t-}) = \lambda(t)dt, \tag{0.1}$$

where dN_t represents the infinitesimal increment of N_t at time t, $\mathcal{F}_{t-}$ represents the past information of the process (what happened before t) and dt is the Lebesgue measure (see [2, Section II.4.1] for more details). Obviously this means that $\lambda(t)$ is random and depends on the past. So one cannot statistically infer $\lambda(t)$ without further assumption on the model. Note however that when we apply this definition to Poisson processes, the independence property in Definition 1 implies that $\lambda(t)$ cannot depend on the past, it is a deterministic function, which corresponds to the intensity s of the Poisson process defined before.

Let us mention two typical examples:

(1) the Aalen multiplicative intensity, i.e.

$$\lambda(t) = Y(t)s(t)dt, \tag{0.2}$$

where $Y(t)$ is a predictable process (i.e. informally, it only depends on the past) that is observed and s is an unknown deterministic function that we want to estimate in an adaptive way. The classical examples covered by this model are right-censoring models,

finite state inhomogeneous Markov processes, etc. We refer to [2] for an extensive list of situations, in particular for biomedical data and survival analysis.

(2) the Hawkes processes, which is defined in the most basic self-exciting model, by

$$\lambda(t) = \nu + \int_{-\infty}^{t^-} h(t-u)dN_u, \tag{0.3}$$

where ν is a positive parameter, h a non negative function with support on $\mathbb{R}^+$ and $\int h < 1$ and where dN_u is the point measure associated to the process. Since $\lambda(t)$ corresponds to (0.1), (0.3) basically means that there is a constant rate ν to have a spontaneous occurrence at t but that also all the previous occurrences influence the apparition of an occurrence at t. For instance an occurrence at u increases the intensity by $h(t-u)$. If the distance $d = t - u$ is favoured, it means that $h(d)$ is really large: having an occurrence at u significantly increases the chance of having an occurrence at t. The intensity given by (0.3) is the most basic case, but variations of it enable us to model self interaction (i.e. also inhibition, which happens when one allows h to take negative values, see Section 2.2.2) and, in the most general case, to model interaction with another type of event.

Hawkes processes have been widely used to model the occurrences of earthquake [56]. The multivariate version can model dependency between action potentials of different neurons in neuroscience [13,39]. It can also be used on genomic data, where this framework was introduced in [21] to model occurrences of events such as positions of genes, promoter sites or words on the DNA sequence. In [12], it has been used to model the positions of transcription regulatory elements in genomic data.

Here the unknown quantity is the couple $s = (\nu, h)$. In biology, very little is known on the function h, except that it should consist in localized spikes at preferred distances/delays corresponding to biological interactions in genomics or neuroscience. The aim of an adaptive procedure is actually to find in practice the localisation of those spikes and their size. Once again, the main ingredient is concentration inequalities.

We begin this review with the link between model selection and Talagrand type inequalities for supremum. Next we deal with more particular methods (such as thresholding or Lasso methods) and their link with Bernstein type inequalities. We finish by a brief overview of the link between tests and concentration inequalities.

Finally, let us just introduce a notation that will avoid tedious explanations. We use in the sequel the notation $\square$ which represents a positive function of the parameters that are written in indices. Each time $\square_\theta$ is written in some equation, one should understand that there exists a positive function of θ such that the equation holds. Therefore the values of $\square_\theta$ may change from line to line and even change in the same equation. When no index appears, $\square$ represents a positive absolute constant.

1. Model Selection and Talagrand type inequalities

The model selection method in its theoretical adaptive presentation has been introduced and developed by Barron, Birgé and Massart [4]. Massart's course [32] in Saint-Flour is one of the best reference on this topic. In particular, Massart emphasizes how concentration inequalities are the fundamental tool to perform model selection. In the next sections, a brief summary of the model selection method is given for Poisson processes, and the corresponding exponential inequality is given. Then the main difficulties arising for other counting processes are emphasized.

1.1. **Poisson framework**

This section is mainly inspired by [41], but we give a simpler version here. We observe a Poisson process N, on $\mathbb{X}$ (say a compact subset of $\mathbb{R}^d$) and we want to estimate its intensity s with respect to the Lebesgue measure, denoted μ. Let T be the Lebesgue measure of $\mathbb{X}$, assumed to be finite.

We work with the $\mathbb{L}_2$-norm defined by

$$\|f\|^2 := \frac{1}{T}\int_{\mathbb{X}} f^2(x)dx,$$

and we define the following least-square contrast

$$\gamma(f) := -\frac{2}{T}\int_{\mathbb{X}} f(x)dN_x + \frac{1}{T}\int_{\mathbb{X}} f^2(x)dx. \tag{1.1}$$

Note that

$$\mathbb{E}(\gamma(f)) = -\frac{2}{T}\int_{\mathbb{X}} f(x)s(x)dx + \frac{1}{T}\int_{\mathbb{X}} f^2(x)dx = \|f-s\|^2 - \|s\|^2,$$

which is minimal when $f = s$. Hence minimizing the least-square contrast should enable us to obtain a good estimator.

Other contrasts may be used. Usually people in point processes theory use MLE ([58], [36], [37]) which corresponds to the minimization of

$$-\frac{2}{T}\int_{\mathbb{X}} \ln(f(x))dN_x + \frac{1}{T}\int_{\mathbb{X}} f(x)dx,$$

but those contrasts are actually more difficult to handle than least-square contrasts for model selection (see [32, Chapter 7] for an extensive comparison of both contrasts in the density setting).

Next the contrast can be minimized on a finite vectorial subspace S of $\mathbb{L}_2$ with orthonormal basis given by $\{\varphi_1, \dots, \varphi_D\}$. The minimization leads to the projection estimator of s

$$\hat{s} = \sum_{i=1}^{D}\left(\int_{\mathbb{X}} \varphi_i(x)\frac{dN_x}{T}\right)\varphi_i. \tag{1.2}$$

Let us study the risk of $\hat{s}$, $\mathbb{E}(\|s-\hat{s}\|^2)$. To do so, let us introduce $\bar{s}$ the orthonormal projection of s on S. This gives

$$\mathbb{E}(\|s-\hat{s}\|^2) = \|s-\bar{s}\|^2 + \frac{1}{T}\sum_{i=1}^{D}\int \varphi_i^2(x)s(x)dx. \tag{1.3}$$

The first term is a bias term, it decreases when S increases whereas the second term, the variance term, increases with the dimension D of S. Obviously finding the best compromise depends on s.

Hence model selection consists in searching for the best S in a family of models (i.e. here, finite vectorial subspaces) $\{S_m, m \in \mathcal{M}_T\}$. To each model S_m let us associate the projection estimator $\hat{s}_m$ and the orthonormal projection of s on S_m, s_m. In a naive approach, the best model we should use, is of course

$$\bar{m} := \text{argmin}_{m\in\mathcal{M}_T}\mathbb{E}(\|s-\hat{s}_m\|^2).$$

This model, $\bar{m}$, is called the oracle. Of course we cannot obtain it without knowing s. One can adapt Mallows' computations [31] to our context and find easily that

$$\bar{m} = \text{argmin}_{m\in\mathcal{M}_T}\left\{\mathbb{E}(\gamma(\hat{s}_m)) + 2\mathbb{E}(\|s_m-\hat{s}_m\|^2)\right\}.$$

It is possible to estimate the previous quantity without bias. Let us denote by $(\varphi_{\lambda,m})_\lambda$ an orthonormal basis of S_m. One obtains the following choice

$$\hat{m} := \text{argmin}_{m\in\mathcal{M}_T}\left\{\gamma(\hat{s}_m) + 2\int\sum_{\lambda}\varphi_{\lambda,m}^2(x)\frac{dN_x}{T^2}\right\}. \tag{1.4}$$

More generally, we consider minimization of the type

$$\hat{m} := \text{argmin}_{m\in\mathcal{M}_T}\left\{\gamma(\hat{s}_m) + \text{pen}(m)\right\} \tag{1.5}$$

and $\tilde{s} = \hat{s}_{\hat{m}}$ is the penalized projection estimator.

We would like to prove that the choice $\hat{m}$ is good, meaning that it can satisfy an oracle inequality in expectation, typically

$$\mathbb{E}(\|\tilde{s} - s\|^2) \leq C\mathbb{E}(\|\hat{s}_{\bar{m}} - s\|^2) = C \inf_{m \in \mathcal{M}_T} \mathbb{E}(\|s - \hat{s}_m\|^2), \tag{1.6}$$

with C an adequate (not too large) multiplicative factor. This would mean that we are able, without knowing $\bar{m}$ to find a model $\hat{m}$ that is performing in essentially the same way. However, we do not obtain (1.6): there is usually a small additive error, which is negligible, and, most importantly, C may grow slowly with T depending on the family of models.

1.1.1. *Histograms*

Let us illustrate this behaviour on the simplest estimators: histograms on an interval (ie $\mathbb{X} = [0, T]$). This example is fundamental to understand what model selection can or cannot do.

Let S_m be a vectorial subspace of $\mathbb{L}_2$ defined by

$$S_m = \left\{ g \quad / \quad g = \sum_{I \in m} a_I \mathbf{1}_I, a_I \in \mathbb{R} \right\},$$

where m is a set of disjoint intervals of $\mathbb{X}$. For histograms, it is actually natural to identify the model S_m and the set m, which is called in the sequel "model" too for sake of simplicity. Let $|m|$ denote the number of intervals in m.

A strategy refers to the choice of the family of models $\mathcal{M}_T$. To avoid any confusion, let $\#\{\mathcal{M}_T\}$ denote the number of models m in $\mathcal{M}_T$. In the sequel, a partition Γ of $[0, T]$ should be understood as a set of disjoint intervals of $[0, T]$ such that their union is the whole interval $[0, T]$. A regular partition is such that all its intervals have the same length. We say that a model m is written on Γ if all the extremities of the intervals in m are also extremities of intervals in Γ. For instance if $\Gamma = \big\{[0, 0.25], (0.25, 0.5], (0.50.75], (0.75, 1]\big\}$ then $\big\{[0, 0.25], (0.25, 1]\big\}$ or $\big\{[0, 0.25], (0.75, 1]\big\}$ are models written on Γ. Now let us give some examples of families $\mathcal{M}_T$. Let J and N be two positive integers.

Nested strategy: Take Γ a dyadic regular partition such that $|\Gamma| = 2^J$. Then take $\mathcal{M}_T$ as the set of all dyadic regular partitions of $[0, T]$ that can be written on Γ. In particular, note that $\#\{\mathcal{M}_T\} = J + 1$. We say that this strategy is nested since for any pair of partitions in this family, one of them is always written on the other one.

Irregular strategy: Assume now that we know that s is piecewise constant on $[0, T]$ but that we do not know where the cuts of the resulting partition are. We can consider Γ a regular partition such that $|\Gamma| = N$ and then consider $\mathcal{M}_T$ the set of all possible partitions written on Γ. In this case $\#\{\mathcal{M}_T\} \simeq 2^N$.

Islands strategy: This last strategy has been especially designed to answer biological questions, i.e. for s having a very localized support. The interval $[0, T]$ is really large and in fact s is non zero on a really smaller interval or a union of really smaller intervals: the resulting model is sparse. We can consider Γ a regular partition such that $|\Gamma| = N$ and then consider $\mathcal{M}_T$ the set of all the subsets of Γ. A typical m corresponds to a vectorial space S_m where the functions g are zero on $[0, T]$ except on some disjoints intervals which look like several "islands". In this case $\#\{\mathcal{M}_T\} = 2^N$.

For all the previous strategies one can prove the following result.

Proposition 1 (Reynaud-Bouret 2003)**.** *Let $\{L_m, m \in \mathcal{M}_T\}$ be a family of positive weights such that $\sum_{m \in \mathcal{M}_T} e^{-L_m |m|} \leq \Sigma$ and assume that $|\Gamma| \leq T(\ln T)^{-2}$ with $T > 2$. For any $c > 1$, if*

$$\text{pen}(m) = \frac{c\tilde{M}|m|}{T}(1 + \sqrt{2\kappa L_m})^2 \text{ with } \tilde{M} = \sup_{I \in \Gamma} \frac{N_I}{\mu(I)},$$

then

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq \square_c \inf_{m\in\mathcal{M}_T}\left[\|s-s_m\|^2 + \frac{M|m|}{T}(1+L_m)\right] + \square_{c,\Sigma,M}\frac{1}{T}, \tag{1.7}$$

where

$$M = \sup_{I\in\Gamma}\frac{\int_I s(x)dx}{\mu(I)}.$$

NB : κ is an absolute constant, namely $\kappa = 6$.

This result is an adapted and simpler version of the one presented in [41], which can be extended to other settings than just histograms.

To shorten mathematical expressions, $\square_c$ and $\square_{c,\Sigma,M}$ have been used even if precise formulas are available. In an asymptotic analysis where T tends to infinity, we consequently need to make c, Σ and M independent of T. However any dependency between $\mathcal{M}_T$ and T is allowed. In this sense, the result of (1.7) (as the ones due to Barron, Birgé and Massart [4]) is non asymptotic with respect to various existing works (such as Mallows' [31]) where the family of models is held fixed whereas T tends to infinity. To obtain (1.7), the fundamental tool is to derive concentration inequalities. Before stating these probabilistic results, let us understand the different behaviours of (1.7) with respect to the different strategies.

Note that for the Nested strategy there exists at most one model m in the family with dimension $|m| = D$ and therefore choosing $L_m = \epsilon > 0$ fixed leads to a quantity Σ independent of T whatever Γ is. We can also remark that $M|m|/T$ is a natural upper bound for the variance term (see (1.3)) and that it is sufficient to assume that s is lower bounded on $\mathbb{X}$ to lower bound the variance term by $r|m|/T$ where $r = \inf_{x\in\mathbb{X}} s(x)$. Therefore the result is an oracle inequality as expected in (1.6) with a true constant C, up to some negligible residual term.

On the other hand, for the Irregular or Islands strategies, there are approximately $(N/D)^D$ models in the family with the same dimension $|m| = D$, therefore one has to take $L_m = \alpha\ln(N)$ or $\alpha\ln(N/D)$ to ensure that Σ will not depend on T whatever Γ is (in particular when the case $N = |\Gamma| = T(\ln T)^{-2}$ is considered). In this case we recover an oracle inequality (see (1.6)) where C is multiple of $\ln(T)$, up to some negligible residual term. This phenomenon is actually unavoidable when considering such complex families of models (i.e. families with complex cardinality: there are more models with the same dimension D than a power of D). Indeed, there exists a minimax lower bound (see Proposition 4 of [41]) that proves the existence of this logarithmic factor. See also [5] and [6] for a more thorough study in the Gaussian setup.

1.1.2. *Concentration inequalities*

The fundamental probabilistic ingredient to show such oracle inequalities is to control the deviations of $\|s_m - \hat{s}_m\|$ which can be written, in the more general setup, as

$$\chi(m) = \sqrt{\sum_\lambda \left(\int_{\mathbb{X}} \varphi_{\lambda,m}(x)\frac{dN_x - s(x)dx}{T}\right)^2},$$

where $(\varphi_{\lambda,m})_\lambda$ is an orthonormal basis of S_m.

In [32], Massart emphasizes the link between Gaussian concentration phenomenon (due to Cirel'son, Ibragimov and Sudakov [14]) and oracle inequalities in the Gaussian setup, but also the link between Talagrand's inequality [53] (and the successive improvements due to Ledoux [30], Massart [33], Klein and Rio [29] or Bousquet [7]) and the density estimation or the classification problem. For Poisson processes, the inequalities of [59] or [24] are not sharp enough to build nice oracle inequalities. Using the infinitely divisible properties of the Poisson process and Ledoux/Massart's approach, one gets the following result

Theorem 1 (Reynaud-Bouret 2003)**.** *Let N be a Poisson process on $\mathbb{X}$ with finite mean measure $\mathfrak{m}$. Let $\{\psi_a, a\in A\}$ be a countable family of functions with values in $[-b;b]$. If*

$$Z = \sup_{a\in A}\int_{\mathbb{X}} \psi_a(x)(dN_x - d\mathfrak{m}_x),$$

then for all $u, \varepsilon > 0$,

$$\mathbb{P}(Z \geq (1+\varepsilon)\mathbb{E}(Z) + 2\sqrt{\kappa v u} + \kappa(\varepsilon) b u) \leq e^{-u},$$

with

$$v = \sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) d\mathfrak{m}_x$$

and $\kappa = 6$, $\kappa(\varepsilon) = 1.25 + 32\varepsilon^{-1}$.

This result, which can be found in [41], has essentially the same flavour as Talagrand's inequality for supremum of empirical processes. The point measure replaces the empirical measure of Talagrand's inequality and the mean measure replaces the expectation. One can also note that the term κ appearing in Theorem 1 is actually the same as the one appearing in the penalty of Proposition 1. The shape of the penalty that is required to obtain an oracle inequality is actually completely related to the shape of this concentration inequality. Indeed it is now easy to obtain an exponential inequality for $\chi(m)$, since

$$\chi(m) = \sup_{f \in S_m, \|f\|=1} \frac{1}{T} \int f(x)(dN_x - s(x)dx).$$

Corollary 1 (Reynaud-Bouret 2003)**.** *Let*

$$M_m = \sup_{f \in S_m, \|f\|=1} \frac{1}{T} \int_{\mathbb{X}} f^2(x) s(x) dx \quad \text{and} \quad B_m = \sup_{f \in S_m, \|f\|=1} \|f\|_\infty.$$

Then for all $u, \varepsilon > 0$,

$$\mathbb{P}\left(\chi(m) \geq (1+\varepsilon)\sqrt{\frac{1}{T}\sum_{\lambda} \int \varphi_{\lambda,m}^2(x) s(x) dx} + \sqrt{\frac{2\kappa M_m u}{T}} + \kappa(\varepsilon)\frac{B_m u}{T}\right) \leq e^{-u}. \tag{1.8}$$

One can see that there are actually two behaviours. When u is small, the behaviour is sub-Gaussian with a variance of the order $M_m/T \leq \|s\|_\infty/T$ which does not grow with the dimension $|m|$ of the model. When u is large, the behaviour is sub-exponential.

There are several improvements of this inequality. For instance, it is possible by restricting oneself to a large event, depending on the model S_m, to privilege the sub-Gaussian behaviour (see Proposition 9 of [41]). This is a classical trick due to Massart, which is easily done once one has a Talagrand type inequality. Using this trick and simplifying a little, the penalty is obtained by keeping the first two terms of (1.8) with $u = L_m$. The fact that $c > 1$ in Proposition 1 is directly connected with the factor $(1+\varepsilon)$ in (1.8).

1.2. **Other counting processes**

Let us present a unified approach for several counting processes. This approach leads to the results of [46] for the Hawkes case in a straightforward way. A slightly different approach has been used in [43] for the Aalen case. Let us recall that the notation s represents the deterministic unknown function appearing in (0.2) for the Aalen setup and that we basically assume that s in this case belongs to

$$\mathbb{L}_2 = \left\{ g \text{ with support in } [0, A] \;\middle/\; \int_0^A g^2 < \infty \right\}.$$

Note that the natural corresponding norm is $\|g\|^2 = \int_0^A g^2$.

For the Hawkes process (see (0.3)), $s = (\nu, h)$ represents the couple where ν is the spontaneous rate of apparition (this is a real number) and h is the interaction function. In this case, we basically assume that s belongs to

$$\mathbb{L}_2 = \left\{ f = (\mu, g) \;\middle/\; g \text{ with support in } (0, A] \text{ and } \int_0^A g^2 < \infty \right\}.$$

In this case, the natural corresponding norm is $\|f\|^2 = \mu^2 + \int_0^A g^2$.

In both cases, the intensity of the process $\lambda(t)$ is of the shape $\Psi_s(t)$, where Ψ is a linear application that transforms any f in the corresponding $\mathbb{L}_2$ space into a predictable process. Indeed for the Aalen case, $\Psi_f(t) = Y_t f(t)$ and in the Hawkes case, $\Psi_f(t) = \mu + \int_{t-A}^{t-} g(t-u)dN_u$. Note that the Poisson process is also of this type with $\Psi_f = f$. Actually, the preliminary forthcoming computations are true for any kind of counting process whose intensity has this linear shape.

Let us observe the counting process N on an interval $[0, T]$ (or $(-A, T]$ for the Hawkes process) and let us define a least-square contrast by

$$\forall f \in \mathbb{L}_2, \quad \gamma(f) = -\frac{2}{T}\int_0^T \Psi_f(t)dN_t + \frac{1}{T}\int_0^T \Psi_f(t)^2 dt. \tag{1.9}$$

Indeed, because of the martingale properties, one easily sees that the compensator of the previous formula at time T is

$$-\frac{2}{T}\int_0^T \Psi_f(t)\Psi_s(t)dt + \frac{1}{T}\int_0^T \Psi_f(t)^2 dt = \frac{1}{T}\int_0^T \Psi_{f-s}(t)^2 dt - \frac{1}{T}\int_0^T \Psi_s(t)^2 dt.$$

Hence the expectation of $\gamma(f)$ is minimal when $\Psi_{f-s}(t) = 0$ for almost every t almost surely. The fact that this implies that $f = s$ depends of course of the process. For the Aalen multiplicative case, this amounts to assume that $\mathbb{E}(Y_t^2) > 0$ for all $t \leq T$ whereas it is more difficult to prove but still true for Hawkes processes when one assumes that h has a bounded support.

We divided by T so that the contrast is exactly the one used for Poisson process, but this division does not change the point were the minimum is reached, so it is not really necessary. Note that for the Poisson process, the division by T is a nice way to introduce asymptotic properties when T tends to infinity. Indeed remark that to derive a true oracle inequality we basically assumed the intensity to be lower bounded. Hence if T grows, the total number of points grows. This vision is still the correct one for Hawkes processes: when T grows, one observes more and more interactions so the estimation should be better. However for the Aalen case, it is not true that the estimation is better when the time T grows. For instance, in the right-censored case, if the hazard rate of the life time of only one patient is estimated, it is not because one observes this patient longer (after his death) that more information is obtained. On the contrary, our estimation will improve when the total number of patients is growing, and this will be true for any kind of *aggregated processes* (i.e. n i.i.d. point processes are observed and one considers their union as process N). In [43], a slightly different least-square contrast was used but it heavily depends on the multiplicative shape of the intensity. Here we only need the linear transformation Ψ.

We can pursue the construction of the projection estimators as before. If S_m is a finite vectorial subspace of $\mathbb{L}_2$ then

$$\hat{s}_m := \operatorname{argmin}_{f \in S_m} \gamma(f). \tag{1.10}$$

Note however that it is not evident to find a closed-form expression for the solution of this minimization. Indeed, with respect to the Poisson case (1.1), on the right hand side of (1.9) appears a random quantity

$$D_T(f) := \frac{1}{T}\int \Psi_f(t)^2 dt$$

which is a random quadratic form on $\mathbb{L}_2$. It happens that in the Poisson case it is the $\mathbb{L}_2$-norm , fact which simplifies several computations.

Next we consider a family of models $\{S_m, m \in \mathcal{M}_T\}$ and we consider again

$$\hat{m} := \operatorname{argmin}_{m \in \mathcal{M}_T} \left\{\gamma(\hat{s}_m) + \operatorname{pen}(m)\right\}, \tag{1.11}$$

and $\tilde{s} = \hat{s}_{\hat{m}}$.

1.2.1. *Concentration inequalities for counting processes*

As we have seen in the Poisson case, the penalty is directly linked to the concentration inequality. Hence, before stating the corresponding oracle inequalities, let us stress the main problems and

results occurring when we deal with general counting processes. Without further details, it is quite obvious to see that the main quantity to control is

$$\chi(m) = \sqrt{\sum_{\lambda} \left(\int_0^T \Psi_{\varphi_{\lambda,m}}(x) \frac{dN_x - \lambda(x)dx}{T} \right)^2} = \sup_{f \in S_m, \|f\|=1} \int_0^T \Psi_f(x) \frac{dN_x - \lambda(x)dx}{T},$$

where $(\varphi_{\lambda,m})_\lambda$ is an orthonormal basis of S_m. In [42], the compensator of a supremum of counting processes is computed. It allows to derive the following result

Theorem 2 (Reynaud-Bouret 2006)**.** *Let $(N_t)_{t\geq 0}$ be a counting process with intensity $\lambda(t)$ assumed to be almost surely integrable on $[0,T]$. Let $\{(H_{a,t})_{t\geq 0}, a \in A\}$ be a countable family of predictable processes and let*

$$\forall t \geq 0, \quad Z_t = \sup_{a\in A} \int_0^t H_{a,s}(dN_s - \lambda(s)ds).$$

Then the compensator $(A_t)_{t\geq 0}$ exists, is non negative et non decreasing and

$$\forall 0 \leq t \leq T, \quad Z_t - A_t = \int_0^t \Delta Z(s)(dN_s - \lambda(s)ds),$$

for a particular predictable process $\Delta Z(s)$ satisfying $\Delta Z(s) \leq \sup_{a\in A} H_{a,s}$.

Moreover, if the $H_{a,s}$'s have values in $[-b,b]$ and if $\int_0^T \sup_{a\in A} H_{a,s}^2 \lambda(s)ds \leq v$ almost surely for some deterministic constants v and b, then for all $u > 0$,

$$\mathbb{P}\left(\sup_{[0,T]} (Z_t - A_t) \geq \sqrt{2vu} + \frac{bu}{3} \right) \leq e^{-u}.$$

This result is a shortened version of Proposition 1 and Theorem 1 of [42]. This result seems more general than Theorem 1 because it deals with general counting processes and as icing on the cake, we obtain an additional supremum on t. But this has a cost. Indeed we can observe that there is an exchange between the supremum and the integral in the definition of v. This cost has already been observed in several frameworks involving dependent structures. For instance the results developed by Wu [59] and Houdré and Privault [24], using martingales techniques, present this exchange. This exchange was also noticed in other dependent setup (see for instance Samson's work on Markov chains [49]).

To understand more precisely what this exchange means, let us apply the previous result to $\chi(m)$.

Corollary 2 (Reynaud-Bouret 2006)**.** *Let*

$$\mathcal{C} = \sum_{\lambda} \int_0^T \frac{\Psi_{\varphi_{\lambda,m}}(x)^2}{T^2} \lambda(x)dx,$$

and assume that $\mathcal{C}$ is bounded by v and $\sum_\lambda \Psi_{\varphi_{\lambda,m}}(x)^2$ is bounded by b for all $x \in [0,T]$. Then, for all $u > 0$,

$$\mathbb{P}\left(\chi(m) \geq \sqrt{\mathcal{C}} + 3\sqrt{2vu} + bu \right) \leq 2e^{-u}.$$

Assume than in our case, one can suppose $(1/T)\int_0^T \Psi_{\varphi_{\lambda,m}}(x)^2 \lambda(x)dx$ bounded by some fixed constant. If we denote by D_m the dimension of S_m then the Gaussian part has a variance of the order D_m/T and grows with the dimension of S_m, whereas it was a constant for the Poisson case. As a consequence, the oracle inequality that we derive using this exponential inequality cannot be as sharp as the one we obtain for Poisson processes in general.

Recently, Baraud [3] proves via chaining arguments a result that supersedes Corollary 2 to some extent. His result actually states, in a more general setup than the one of counting processes, that one can obtain a Gaussian part with dimension-free variance at the cost of a larger constant term

(i.e. the concentration phenomena in his case is not around $\sqrt{\mathcal{C}}$ but around something larger). In good cases (special choices of $u \simeq D_m$ and "nice" counting processes), it may happen that one recovers the order of magnitude of the Poisson case instead of the present deteriorate rate. However the fact that the leading term which replaces $\sqrt{\mathcal{C}}$ is not the expectation of $\chi(m)$ but some thing much larger, makes this kind of inequality quite unsuitable for practical purpose.

1.2.2. *Oracle inequalities*

It is quite difficult to write a general oracle inequality, because it heavily depends on the norm one considers. The natural norm we would like to consider is $D_T(f)$. But $D_T(f)$ is a random quadratic form and not strictly speaking a norm: it may eventually be null for some non zero f. Of course this function f would have to be random and very peculiar. It is easier to understand it in the Aalen case, even if the same phenomenon applies for Hawkes processes. If $Y_t = 0$ on some subinterval of $[0, T]$, then a function f which is non zero on this random interval is a solution. Assuming that $\mathbb{E}(Y_t^2) > 0$ on the whole interval $[0, T]$ does not prevent the random variable to be null eventually. For the Aalen case, one has to restrict oneself to the event $\{Y_t$ bounded from below on $[0, T]\}$. More generally we will have to restrict oneself at least to the event

$$\mathcal{E} = \left\{\forall m \in \mathcal{M}_T, \quad \forall f \in S_m, \quad r^2\|f\|^2 \leq D_T^2(f) \leq R^2\|f\|^2\right\}, \tag{1.12}$$

for some fixed constants r and R, with $\|f\|$ the natural norm on $\mathbb{L}_2$. But then of course the resulting oracle inequality (1.6) cannot hold in expectation on the whole probability space. To do so, among other technicalities, one obviously needs to control $\mathbb{P}(\mathcal{E}^c)$ and this is basically not related to the martingale structure of the counting process but to some additional properties.

Aalen multiplicative intensity. For the Aalen case, the additional properties may come from the aggregated case. Let us just give a brief summary of the type of oracle inequalities that can be found in [43].

- If one uses histograms, and if, among other technical assumptions, one assumes that N is a bounded aggregated process, then a result strictly equivalent to Theorem 1 is available, since Talagrand's inequality can be used on the aggregated process.
- If one uses random models, with known orthonormal basis for $D_T(f)$, then one is forced to use the exponential inequality of Corollary 2.
  - Hence the oracle inequality is limited to not too complex families of models. One model per dimension is the basic case, for which the penalty should be $\text{pen}(m) = cD_m/T$ for some large enough constant c.
  - the oracle inequality is stated as follows

$$\mathbb{E}(D_T(s - \tilde{s})\mathbf{1}_{\mathcal{E}}) \leq \square_{c,s} \inf_{m \in \mathcal{M}_T} \left[\mathbb{E}(D_T(s - \hat{s}_m)) + \frac{D_m}{T}\right] + \square_{c,s}\frac{1}{T}.$$

  - One can control $\mathcal{E}$ if one assumes again the process to be aggregated.

Note that the approach based on least-square contrasts for aggregated counting processes has been developed and widened by Brunel and Comte (and co-authors) in a succession of papers, in particular under various type of censoring (see for instance [10] and [11]).

Hawkes processes. For the Hawkes process, one cannot use that N is an aggregated process any more and that the individual processes are more or less bounded. It is true that the Hawkes process is infinitely divisible but it is typically unbounded, the number of points per interval being sensibly larger than a Poisson variable (exponential moments exist but not of any order). The concentration for infinite divisible variables developed in [26] cannot be applied directly to the resulting $\chi(m)$. Indeed $\chi(m)$ can be viewed as the norm of a random vector of infinitely divisible variables, but the structure of Ψ does not allow those variables to be independent.

However, one still needs to control the event $\mathcal{E}$. Actually, the Hawkes process has ergodic properties that show that $D_T(f)$ tends to a true norm on $\mathbb{L}_2$ when T tends to infinity. In [44], two main types of exponential inequalities for Hawkes processes have been derived. First, a control of the number of points of the Hawkes process per interval is given. Second and most importantly, a

non asymptotic control of the rate of convergence in the ergodic theorem is inferred. The results of [44] imply the following result:

Lemma 1 (Reynaud-Bouret Roy 2007)**.** *Let $(N_t)_{t\in\mathbb{R}}$ be a stationary Hawkes process, with intensity given by $\lambda(t)=\Psi_s(t)$ with $s=(\nu,h)$ in $\mathbb{L}_2$ and positive h. Note that the definition of $\mathbb{L}_2$ implies that the interaction function h has a bounded support included in $(0,A]$. Let g be a function of the points of $(N_t)_{t\in\mathbb{R}}$ lying in $[-A,0)$, with values in $[-B,B]$ and zero mean. Let $(\theta_t)_{t\in\mathbb{R}}$ be the flow induced by $(N_t)_{t\in\mathbb{R}}$ i.e. $g\circ\theta_t$ is the same function as before, but now the points are lying in $[-A+t,t)$. Then there exists a positive constant $T_0(p,A)$ depending on $p=\int h$ and A, such that for all $T\geq T_0(p,A)$*

$$\mathbb{P}\left(\left|\frac{1}{T}\int_0^T g\circ\theta_t dt\right|\geq 2\sqrt{\frac{c_1\mathrm{Var}(g)A\log(T)^2}{T(p-\log p-1)}}+\frac{c_2BA\log(T)^2}{T(p-\log p-1)}\right)\leq\frac{\square_{\nu,p}}{T^3},$$

where c_1 and c_2 are absolute constants.

The key of the proof is the behaviour of the process not in terms of martingale, as before, but in terms of branching process, when h is non negative (see [23]). This lemma is the main probabilistic tool for controlling $\mathcal{E}$. A more precise statement and variations of Lemma 1 can be found in [44].

Now we can consider as models S_m, sets of couples (μ,g) where μ is any real number and where g is a piecewise constant function on a set m of intervals of $(0,A]$. All the strategies, i.e. families of possible m's, that have been described as histograms strategies for Poisson processes apply here.

The only remaining problem is that we need to control $\tilde{s}$ on $\mathcal{E}^c$, which can be done theoretically via clipping. Let us define for all real numbers $H>0$, $\eta>\rho>0$, $1>P>0$, the following subset of $\mathbb{L}_2$:

$$\mathcal{L}_{H,P}^{\eta,\rho}=\left\{f=(\mu,g)\in\mathbb{L}_2\Big/\quad \mu\in[\rho,\eta],\quad g(.)\in[0,H]\text{ and }\int_0^A g(u)du\leq P\right\},$$

and let us assume that we know that s belongs to this set. Recall that the penalized projection estimator $\tilde{s}=(\tilde{\nu},\tilde{h})$ is given by (1.11). Then, under the previous assumption, it is natural to consider the clipped penalized projection estimator, $\bar{s}=(\bar{\nu},\bar{h})$, given, for all positive t, by

$$\left\{\begin{array}{rcl}\bar{\nu} & = & \left\{\begin{array}{lll}\tilde{\nu} & \text{if} & \rho\leq\tilde{\nu}\leq\eta,\\ \rho & \text{if} & \tilde{\nu}<\rho,\\ \eta & \text{if} & \tilde{\nu}>\eta,\end{array}\right.\\ \bar{h}(t) & = & \left\{\begin{array}{lll}\tilde{h}(t) & \text{if} & 0\leq\tilde{h}(t)\leq H,\\ 0 & \text{if} & \tilde{h}(t)<0,\\ H & \text{if} & \tilde{h}(t)>H.\end{array}\right.\end{array}\right. \tag{1.13}$$

Theorem 3 (Reynaud-Bouret Schbath 2010)**.** *Let $(N_t)_{t\in\mathbb{R}}$ be a Hawkes' process with intensity $\Psi_s(.)$. Assume that we know that s belongs to $\mathcal{L}_{H,P}^{\eta,\rho}$. Moreover assume that all the models in $\mathcal{M}_T$, i.e. possible sets m of intervals, are written on Γ, a regular partition of $(0,A]$ such that*

$$|\Gamma|\leq\frac{\sqrt{T}}{(\log T)^3}. \tag{1.14}$$

Let $Q>1$. Then there exists a positive constant κ depending on η,ρ,P,A,H such that if

$$\forall m\in\mathcal{M}_T,\quad \mathrm{pen}(m)=\kappa Q\big(|m|+1\big)\frac{\log(T)^2}{T}, \tag{1.15}$$

then

$$\mathbb{E}\big(\|\bar{s}-s\|\big)^2\leq\square_{\eta,\rho,P,A,H}\inf_{m\in\mathcal{M}_T}\left[\|s-s_m\|^2+\big(|m|+1\big)\frac{\log(T)^2}{T}\right]+\square_{\eta,\rho,P,A,H}\frac{\#\{\mathcal{M}_T\}}{T^Q},$$

where s_m is the orthogonal projection of s on S_m.

One can compare Proposition 1 and Theorem 3. First $|\Gamma|$ should be smaller for the Hawkes process: this comes basically from the control of $\mathcal{E}$, which was unnecessary for the Poisson process since $D_T(f)$ was deterministic in this case. Next, weights - the L_m's - were appearing in the Poisson case: they are replaced here by the factor $Q\log(T)$. Actually the concentration we used (Corollary 2) is not sharp enough to use weights as precisely as in Proposition 1. Indeed since the dimension appears in the variance term in Corollary 2, one needs basically to take $u = Q\ln(T)$ to obtain a deviation of order $\sqrt{D_m/T}$ up to some logarithmic term. On the contrary, the variance does not depend on the dimension in Corollary 1 for the Poisson case and one can take $u = L_m D_m$.

In addition, the penalty has an extra $\log(T)$ factor which comes from the fact that the intensity $\lambda(t)$ is unbounded: $\lambda(t)$ behaves basically as the number of points in an interval of length A, quantity for which tail estimates are available in [44]. As we want to control it on the whole interval $[0, T]$ we lose an extra logarithmic factor.

Hence, if the results look similar between Proposition 1 and Theorem 3, up to logarithmic factors, we can note the following main differences. For the Nested strategy, the constraint $|\Gamma| << \sqrt{T}$ slightly limits the size of the family and the penalty is a little bit larger with extra-logarithmic factors. For the Irregular and Islands strategies, the limitation in size is much more drastic since one needs $|\Gamma| \leq \square \log(T)$ to ensure that the residual term $\#\mathcal{M}_T/T^Q$ is not exploding with T.

1.3. **Main drawbacks of model selection**

There are two main criticisms to the previous method.

First, the computational cost of such general model selection method may be too high to be of real interest in practice. For instance, when one consider the Islands family of models with a regular partition Γ of length, say 25, one needs to compute 2^{25} models, reaching very rapidly the limit memory of any existing computer. There are several ways to avoid such a problem. Either there are algorithmic simplifications, using either thresholding rules or dynamic programming, where one can avoid to compute all the models before selecting one. But those kind of simplifications are not always possible, in particular for Hawkes processes. Or one can use a convex criteria, typically a Lasso penalty, to obtain an implementable algorithm on classical data size.

The second main drawback is due to the concentration itself. Indeed, comparing the previous results, one clearly sees that the better the concentration, the better the penalty and the more complex the family of models. But even in the nicest case, i.e. the Poisson case, the term v in Theorem 1 is not the variance of the supremum Z but the supremum of the variances. This fact implies that in Proposition 1, the penalty term involves $\tilde{M}|m|/T$ as an estimate not of the variance of $\hat{s}_m$, but of an upper bound of the variance, namely $M|m|/T$ (see also (1.3) and the ideal Mallows penalty in (1.4)). Theoretically speaking, this does not change the order of magnitude of the oracle inequality as we have seen in Proposition 1 but in practice if the function has very localised spikes (as we expect for some data), using this upper bound instead of the true variance, makes us lose the right order of magnitude and the resulting estimate will have a very poor behavior. Of course when dealing with other counting processes the exchange between integral and supremum in v, as noticed before, is not only deteriorating the practical performance, it is also deteriorating the general theoretical performance of the estimate.

2. Thresholding, Lasso and Bernstein type inequalities

If we want sharper concentration inequalities, we have to focus on simpler concentration inequalities involving not a supremum of processes but only one process at a time. So we turn toward Bernstein type inequalities [32], where the quantity v is indeed the variance of the quantity whose deviation is computed. Consequently there are two main issues. First we need other adaptive statistical methods, which can be proved to satisfy oracle inequalities if we can only provide Bernstein type inequalities. These methods are less general than model selection methods, but can be calibrated in a more precise way. Next, we need to provide Bernstein type inequalities where the variance v can be replaced by a data-driven quantity of the same order, so that the resulting method is completely data-driven.

2.1. **Thresholding and Poisson processes**

Thresholding methods can be viewed as a very particular example of model selection (see [32]). But actually because they are much simpler, they can reached adaptive properties that cannot be proved for other methods. In particular, there exists the following very general result, which can also been considered as an oracle inequality (see [45]).

Theorem 4 (Reynaud-Bouret, Rivoirard 2010)**.** *To estimate a sequence $\beta = (\beta_\lambda)_{\lambda\in\Lambda}$ such that $\|\beta\|_{\ell_2} < \infty$, two observable sequences $(\hat\beta_\lambda)_{\lambda\in\Gamma}$ and $(\eta_\lambda)_{\lambda\in\Gamma}$ are given, where $\Gamma \subset \Lambda$.*
Consider $\tilde\beta = (\hat\beta_\lambda \mathbf{1}_{|\hat\beta_\lambda|\geq\eta_\lambda}\mathbf{1}_{\lambda\in\Gamma})_{\lambda\in\Lambda}$ the thresholding estimate of the sequence β.
Let $\epsilon > 0$ be fixed. If there exists $(F_\lambda)_{\lambda\in\Gamma}$ and $\kappa \in [0,1[$, $\omega \in [0,1]$, $\zeta > 0$ such that

(A1) *For all λ in Γ, $\mathbb{P}\left(|\hat\beta_\lambda - \beta_\lambda| > \kappa\eta_\lambda\right) \leq \omega$.*

(A2) *There exists $1 < a, b < \infty$ with $\frac{1}{a} + \frac{1}{b} = 1$ and $G > 0$ such that $\lambda \in \Gamma$,*

$$\left(\mathbb{E}\left[|\hat\beta_\lambda - \beta_\lambda|^{2a}\right]\right)^{\frac{1}{a}} \leq G\max\left(F_\lambda,\ F_\lambda^{\frac{1}{a}}\epsilon^{\frac{1}{b}}\right).$$

(A3) *There exists τ such that for all λ in Γ / $F_\lambda < \tau\epsilon$, $\mathbb{P}\left(|\hat\beta_\lambda - \beta_\lambda| > \kappa\eta_\lambda,\ |\hat\beta_\lambda| > \eta_\lambda\right) \leq F_\lambda\zeta$.*

Then

$$\mathbb{E}\|\tilde\beta - \beta\|_{\ell_2}^2 \leq \square_\kappa \mathbb{E}\inf_{m\subset\Gamma}\left\{\sum_{\lambda\notin m}\beta_\lambda^2 + \sum_{\lambda\in m}(\hat\beta_\lambda - \beta_\lambda)^2 + \sum_{\lambda\in m}\eta_\lambda^2\right\} + \square_{G,\kappa,\tau,\omega,\epsilon,\zeta}\sum_{\lambda\in\Gamma}F_\lambda.$$

The main two ingredients are (A1), a control of the difference between the basic estimate $\hat\beta_\lambda$ and its target β_λ by the threshold η_λ and (A2), a Rosenthal type inequality. Assumption (A3) is a very technical assumption that is usually fulfilled.

Let us apply this theorem to Poisson processes. Assuming that s in an $\mathbb{L}_2$ function, we decompose the intensity s on the Haar basis, i.e.

$$s = \sum_{j\geq -1}\sum_{k\in\mathbb{Z}}\beta_{j,k}\varphi_{j,k}, \tag{2.1}$$

where for any $j \geq -1$ and any $k \in \mathbb{Z}$,

$$\beta_{j,k} = \int s(x)\varphi_{j,k}(x)dx,$$

with for any $x \in \mathbb{R}$,

$$\varphi_{-1,k}(x) = \mathbf{1}_{k\leq x<k+1}, \quad \text{and for all j} \geq 0, \varphi_{j,k}(x) = 2^{\frac{j}{2}}\left(\mathbf{1}_{0\leq 2^j x-k<1/2} - \mathbf{1}_{1/2\leq 2^j x-k<1}\right).$$

Hence we identify $\lambda = (j,k)$, $\beta_\lambda = \beta_{j,k}$, $\Lambda = \{(j,k), k \in \mathbb{Z}, j \geq -1\}$ and $\Gamma = \{(j,k), k \in \mathbb{Z}, j \leq j_0\}$, where j_0 is some largest resolution level. In the Poisson setup , $\hat\beta_\lambda$ is easy to find, this is

$$\hat\beta_\lambda = \int \varphi_\lambda(x)dN_x.$$

For a given thresholding estimator $\tilde\beta$, as in Theorem 4, of the sequence β, one can associate $\tilde s$, the classical thresholding estimator of s, by

$$\tilde s = \sum_{\lambda\in\Gamma}\tilde\beta_\lambda\varphi_\lambda.$$

Therefore Theorem 4 gives conditions on the threshold η_λ such that $\tilde s$ fulfills classical oracle inequalities. We refer to [45] for the solution of (A2) and (A3) and we just focus on (A1). This

is where Bernstein type inequalities come into play. Indeed, there is a well-known result, namely Campbell's Theorem [28], which states that for all measurable function f,

$$\mathbb{E}\left[\exp\left(\int f(x)dN_x\right)\right] = \exp\left(\int (e^{f(x)}-1)d\mathfrak{m}_x\right),$$

as soon as the right hand side is finite. Inverting this formula as for Bernstein's inequality (see [32] for instance), leads to this Bernstein type inequality (see also [41]): for all positive u,

$$\mathbb{P}\left(\int f(x)(dN_x - d\mathfrak{m}_x) \geq \sqrt{2u\int f^2(x)d\mathfrak{m}_x} + \frac{1}{3}\|f\|_\infty u\right) \leq \exp(-u), \tag{2.2}$$

as soon as $\int f^2(x)d\mathfrak{m}_x$ and $\|f\|_\infty$ are finite.

But if we combine directly this result with (A1), we obtain a threshold that is not observable, indeed of the form

$$\eta_\lambda = \sqrt{2u\int \varphi_\lambda^2 s(x)dx} + \frac{1}{3}\|\varphi_\lambda\|_\infty u,$$

for some u of typical logarithmic order (see [45] for more details). More explicitly, a classical Bernstein type inequality allows us to use the exact variance in the threshold (instead of an upper bound of the variance as in the model selection penalty) but since the variance is unknown it cannot be used in practice. But it is easy to derive this Lemma from the previous result.

Lemma 2 (Reynaud-Bouret, Rivoirard 2010)**.** *For any $u > 0$,*

$$\mathbb{P}\left(\left|\int f(x)(dN_x - d\mathfrak{m}_x)\right| \geq \sqrt{2u\breve{V}_f(u)} + \frac{\|f\|_\infty u}{3}\right) \leq 3e^{-u}. \tag{2.3}$$

where

$$\breve{V}_f(u) = \hat{V}_f + \sqrt{2\hat{V}_f\|f\|_\infty^2 u} + 3\|f\|_\infty^2 u,$$

with $\hat{V}_f = \int f^2(x)dN_x$.

This means that we can take

$$\eta_\lambda = \sqrt{2u\breve{V}_{\varphi_\lambda}(u)} + \frac{1}{3}\|\varphi_\lambda\|_\infty u, \tag{2.4}$$

which is observable as a threshold, for some well-chosen u of logarithmic order.

In [45], using this main idea, several things have been proved.

- A classical oracle inequality on $\tilde{s}$ exists and this even if s is unbounded, or has no finite support. So in this sense, it supersedes Proposition 1 or its generalisations of [41].
- Because now the variance "v" factor is now $\breve{V}_{\psi_\lambda}(u)$, which is, up to negligible term, an unbiased estimate of the variance of $\hat{\beta}_\lambda$, one can prove precise lower bounds for the threshold in the spirit of [6] for model selection penalty in the Gaussian setup. More precisely, if the thresholds η_λ's are given by (2.4) with $u = \gamma\ln(n)$ in an aggregated framework, and if $\gamma < 1$, then the rate of convergence of $\tilde{s}$ is worse than when $\gamma > 1$ for true functions s as simple as $\mathbf{1}_{[0,1]}$.
- Because the fluctuations of $\hat{\beta}_\lambda$ around its mean are now very well approximated by the threshold and not just largely upper bounded by it, the practical performance of the corresponding thresholding estimator is very good and it has been used in neuroscience [48] for instance.

This kind of approach by thresholding rules coupled with sharp observable thresholds resulting from Bernstein type inequalities has also been applied to other frameworks: density estimation [47] and Poissonian interaction models [50].

2.2. **Lasso criterion and other counting processes**

For more general counting processes, under the assumption of linearity of the intensity, the same kind of considerations can be done, except that the adaptive method that can be used is not the thresholding method but the Lasso method, since there is no easy access to unbiased estimate of $\hat{\beta}_\lambda$ in general.

We follow [22] except that the framework is restricted to univariate counting processes, for sake of simplicity. We consider again general linear predictable intensity of the type Ψ_s for s in some Hilbert $\mathbb{L}_2$ space. More precisely, following [22], we consider a dictionary Φ of $\mathbb{L}_2$ and denote for any $a \in \mathbb{R}^\Phi$

$$f_a = \sum_{\varphi\in\Phi} a_\varphi \varphi.$$

Given positive weights $(d_\varphi)_{\varphi\in\Phi}$, the Lasso estimate of s is $\hat{s} := f_{\hat{a}}$ where $\hat{a}$ is a minimizer of the following ℓ_1-penalized least-square contrast (see also (1.9)):

$$\begin{aligned}\hat{a} &:= \operatorname{argmin}_{a\in\mathbb{R}^\Phi}\left\{\gamma(f_a) + \sum_{\varphi\in\Phi} d_\varphi |a_\varphi|\right\} \\ &= \operatorname{argmin}_{a\in\mathbb{R}^\Phi}\{-2a'b + a'Ga + 2d'|a|\},\end{aligned}$$

where for any φ and $\tilde{\varphi}$,

$$b_\varphi = \int_0^T \Psi_\varphi(t) dN_t \quad G_{\varphi,\tilde{\varphi}} = \int_0^T \Psi_\varphi(t)\Psi_{\tilde{\varphi}}(t)dt,$$

are observable data-driven quantities and where a' denotes the transpose of the vector a.

Because the criterion is convex, the computation cost of the minimization is reasonable. Moreover a general oracle inequality in the spirit of Theorem 4 can be found in [22], except that it only holds in probability. The main condition, which replaces (A1) is that the probability of the following event is small:

$$\left\{\forall\varphi\in\Phi, \quad \left|\int_0^T \Psi_\varphi(dN_t - \lambda(t)dt)\right| \le d_\varphi\right\}.$$

Once again the main point is a Bernstein type inequality in the spirit of Lemma 2.3. Let us give here a simple version, the multivariate version being in [22].

Theorem 5. *Let $(N_t)_{t\ge0}$ be a counting process with intensity $\lambda(t)$ assumed to be almost surely integrable on $[0,T]$. Let $(H_s)_{s\ge0}$ be a predictable process and $M_t = \int_0^t H_s(dN_s - \lambda(s)ds)$. Let $b>0$ and $v > w > 0$ such that for all $\xi \in (0,3)$, for all t,*

$$\int_0^t \exp(\xi H_s/b)\lambda(s)ds < \infty \text{ a.s. and } \int_0^t \exp(\xi H_s^2/b^2)\lambda(s)ds < \infty \text{ a.s.} < \infty \text{ a.s.} \tag{2.5}$$

For all $x, \mu > 0$ such that $\mu > \phi(\mu)$, let

$$\hat{V}_\tau^\mu = \frac{\mu}{\mu - \phi(\mu)}\int_0^\tau H_s^2 dN_s + \frac{b^2 x}{\mu - \phi(\mu)},$$

where $\phi(u) = \exp(u) - u - 1$.

Then for every stopping time τ and every $\varepsilon > 0$

$$\mathbb{P}\left(M_\tau \ge \sqrt{2(1+\varepsilon)\hat{V}_\tau^\mu x} + bx/3, \quad w \le \hat{V}_\tau^\mu \le v \text{ and } \sup_{s\in[0,\tau]}|H_s| \le b\right) \le 2\frac{\log(v/w)}{\log(1+\varepsilon)}e^{-x}.$$

This result is based on the exponential martingale for counting processes, which has been used for a long time in the context of counting process theory. See for instance [9], [52] or [55]. This basically gives a concentration inequality taking the following form, which is the general counting process equivalent of (2.2): for any $x > 0$,

$$\mathbb{P}\left(M_\tau \geq \sqrt{2\rho x} + \frac{bx}{3} \text{ and } \int_0^\tau H_s^2 \lambda(s) ds \leq \rho \text{ and } \sup_{s\in[0,\tau]} |H_s| \leq b \right) \leq e^{-x}. \tag{2.6}$$

In (2.6), ρ is a deterministic upper bound of $v = \int_0^\tau H_s^2 \lambda(s) ds$, the bracket of the martingale, and consequently the martingale equivalent of the variance term. Moreover b is a deterministic upper bound of $\sup_{s\in[0,\tau]} |H_s|$. The leading term for moderate values of x and τ large enough is consequently $\sqrt{2\rho x}$ where the constant $\sqrt{2}$ is not improvable since this coincides with the rate of the central limit theorem for martingales. However in the central limit theorem for martingales, it is assumed that v tends to a deterministic value, which is the asymptotic variance (once everything is correctly renormalized). So if the $\sqrt{2}$ is not improvable, it is likely that a fixed deterministic value which upper bounds v constitutes a loss. In this sense, Theorem 5 improves the bound and consists in plugging the estimate $\hat{v} = \int_0^\tau H_s^2 dN_s$ instead of a non sharp deterministic upper bound of v. Two small losses have to be underlined: we are not able to recover exactly the constant $\sqrt{2}$ but any value strictly larger than $\sqrt{2}$, as close as we want to $\sqrt{2}$ and we lose some additive terms depending on b that are negligible for moderate values of x.

Hence we are able to obtain a much sharper concentration inequality for one single integral than for a supremum (see Theorem 2), with respect to the discussions of Section 1.3. Of course it is a simpler framework (only one process), but this inequality helps us to furnish a fully data-driven choice of the weights d_φ, as in (2.4), leading to a fully calibrated method for general counting processes (see [22] for more details). The counterpart is that we cannot do general model selection with general models S_m but are forced to search the estimate within a dictionary (and under particular conditions on the dictionary). The resulting method is very well calibrated as shown on the simulations of [22] and has been successfully applied on real neuronal data [48].

Lasso (and similar) methods for particular counting processes such as Cox model or multiplicative Aalen intensity have also been derived for instance in [8] or [19].

3. Test and U-statistics

We conclude this review with a very small section dedicated to adaptive testing. It is much more difficult to present tests in unified way. Indeed depending on whether one tests goodness-of-fit [27], homogeneity [17], two-sample problems [18] or independence [51], the test statistics may have very different shapes and therefore it is difficult to point out which bounds may be useful. So let us just illustrate where exponential inequalities are useful when dealing with testing on the particular framework of [17].

We observe a Poisson process N with unknown intensity $s(x)$ wrt ndx on $[0,1]$, where n is a positive integer, to simplify notations. Here n replaces T and tends to infinity for asymptotic purposes, this setup being equivalent to the aggregation of n i.i.d. Poisson processes on $[0,1]$ with intensity s wrt dx.

We assume that $\|s\|_\infty < \infty$ and that one can again decompose s on the Haar basis. Since we are on $[0,1]$, this amounts to

$$s = \beta_{-1,0}\varphi_{-1,0} + \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \beta_{j,k}\varphi_{j,k}.$$

We want to test H_0: "s is constant" (i.e. N is a homogeneous Poisson process) against H_1: "s is not constant". Adaptive testing procedures consist in designing a test that is powerful for a wide class of possible spaces as alternatives.

Let us just understand where the concentration inequalities may be useful on one finite vectorial subspace. Let $m \subset \{(j,k), j \geq 0, k = 0, ..., 2^j - 1\}$ and consider the model $S_m = \text{Span}(\varphi_{-1,0}, \varphi_\lambda, \lambda \in$

m) with dimension D_m. When we want to test the homogeneity, we actually want to reject when the distance between s and $S_0 = \mathrm{Span}(\varphi_{-1,0})$ is too large. This distance can be estimated and the estimate may be used as test statistic. This idea is very old. It has been introduced in the Poisson setting by Watson [57]. The procedure is consequently decomposed as follows:

(1) We approximate $d(s, S_0)^2$ by $\sum_{\lambda\in m} \beta_\lambda^2$.
(2) We unbiasly estimate it by $T_m = \sum_{\lambda\in m} T_\lambda$ with

$$T_\lambda = \hat{\beta}_\lambda^2 - \frac{1}{n^2}\int \varphi_\lambda^2 dN.$$

(3) Under H_0 the distribution of T_m given that $N_{[0,1]} = K$ is free of s, so there exists $t_{m,\alpha}^{(K)}$, the $1-\alpha$ quantile of the conditional distribution of T_m, such that

$$\mathbb{P}(T_m > t_{m,\alpha}^{(K)} | N_{[0,1]} = K) \leq \alpha.$$

(4) We consequently reject H_0 when $T_m > t_{m,\alpha}^{(N_{[0,1]})}$.

Hence concentration inequalities are not helpful for calibrating the test (this is done at step 3 by using the exact quantiles). But they are helpful to find bounds on the separation distance, i.e. to answer the question "under H_1, how far from S_0 should s be to obtain $\mathbb{P}(\text{accept } \mathrm{H}_0) \leq \beta$?" for some fixed error $\beta \in [0, 1]$.

If there exists a quantity $A_{m,\alpha,\beta}$ such that

$$\mathbb{P}(t_{m,\alpha}^{(N_{[0,1]})} \geq A_{m,\alpha,\beta}) \leq \beta/3, \tag{3.1}$$

and if

$$d^2(s, S_0) \geq \|s - s_m\|^2 + \square_{\beta,\|s\|_\infty} \frac{\sqrt{D_m}}{n} + A_{m,\alpha,\beta},$$

then, under suitable assumptions, it is easy to prove that the error of second kind is less than β (see the precise version with slightly different notations in Theorem 4 of [17]). However the presence of $A_{m,\alpha,\beta}$ is crucial.

One can prove, using exponential inequalities for degenerate U-statistics of order 2 - which is the case for T_m - that

$$A_{m,\alpha,\beta} = \square_{\beta,\|s\|_\infty} \left[\frac{\sqrt{D_m \log(\alpha^{-1})}}{n} + \frac{\log(\alpha^{-1})}{n} + \frac{E_m \log^2(\alpha^{-1})}{n^2} \right],$$

satisfies (3.1) where $E_m = \sum_{j/(j,k)\in m} 2^j$ may be much larger than D_m.

The logarithmic dependency of $A_{m,\alpha,\beta}$ in terms of α is fundamental, because when we want to turn those tests into adaptive tests, we have to consider a collection of tests (i.e. various choices for m) and we have to roughly divide the level α by the number of tests in order to guarantee a fixed level α. Hence the lighter the dependence in α, the larger the number of tests that can be used together. This explains the need for concentration inequalities for U-statistics, when dealing with tests of homogeneity, but also the need for control of Rademacher chaos for the two-sample problem [18], etc. Each time, concentration inequalities will give the behaviour of the quantiles of the test statistics in terms of α. Hence, in testing problems, concentration inequalities do not need to have sharp constants since they are not used for practice but only for theory.

On this particular example, for U-statistics, note that exponential inequalities are described in the book of de la Peña and Giné [16]. These upper bounds have been improved but still with unknown constants by Giné, Latala and Zinn [20]. In [25], precise constants in those formula are derived by combining Talagrand's inequality and martingale properties for degenerate U-statistics of order 2. This also applies to Poisson processes since one can replace Talagrand's inequality by Theorem 1 (see [25] for more details). Let us also mention [1], which involves degenerate U-statistics of any order for independent variables but also for processes with independent increments.

Conclusion

Adaptive statistics can exist only because of specific probabilistic inequalities and this is true also in the setup of counting processes. The present review of such inequalities is of course not exhaustive, but I hope that the concentration inequalities that have been used in the works presented here are general enough to be of interest for other researchers, who are looking for specific exponential inequalities for counting processes.

References

[1] Adamczak, R. *Moment Inequalities for U-statistics.* Ann. Probab. **34** (6), 2288–2314 (2006).

[2] Andersen, P. K., Borgan, Ø., Gill, R. D., Keiding, N. *Statistical models based on counting processes.* Springer Series in Statistics (1993).

[3] Baraud, Y. *A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression.* Bernoulli, **16**(4), 1064–1085 (2010).

[4] Barron, A., Birgé, L., Massart, M. *Risk bounds for model selection via penalization.* Probab. Theory Related Fields, **113**(3), 301–413 (1999).

[5] Birgé, L., Massart, P. *Gaussian model selection.* J. Eur. Math. Soc. **3**(3), 203–268 (2001).

[6] Birgé, L. Massart, P. *Minimal penalties for Gaussian model selection.* P.T.R.F. **138**(1-2), 33–73 (2007).

[7] Bousquet, O. *A Bennett concentration inequality and its application to suprema of empirical processes.* C. R. Acad. Sci. Paris, Ser. I **334**, 495–500 (2002).

[8] Bradic, J., Fan, J. Jiang, J. *Regularization for Cox's Proportional Hazards Model with NP-Dimensionality.* Annals of Statistics, **39**(6), 3092–3120 (2011).

[9] Brémaud, P. *Point processes and queues.* Springer Series in Statistics (1981).

[10] Brunel, E., Comte, F. *Penalized contrast estimation of density and hazard rate with censored data.* Sankhya **67**(3), 441–475 (2005).

[11] Brunel, E., Comte, F. *Adaptive estimation of hazard rate with censored data.* Communications in Statistics, Theory and methods **37**(8), 1284–1305 (2008).

[12] Carstensen, L., Sandelin, A., Winther, O., Hansen, N.R., *Multivariate Hawkes process models of the occurrence of regulatory elements.* BMC Bioinformatics (2010).

[13] Chornoboy, E.S., Schramm, L.P., Karr, A.F. *Maximum likelihood identification of neural point process systems.* Biological Cybernetics, **59**, 265–275 (1988).

[14] Cirel'son, B.S., Ibragimov, I.A., Sudakov, V.N. *Norms of gaussian sample functions.* Proc. 3rd Japan-USSR Symp. Probab. Theory, Taschkent 1975, LNM **550**, 20–41 (1976).

[15] Daley, D.J., Vere-Jones, D. *An introduction to the theory of point processes.* Springer series in statistics Volume I (2005).

[16] de la Peña, V, Giné, E. *Decoupling : from dependence to independence.* Springer series in statistics (1999).

[17] Fromont, M., Laurent, B., Reynaud-Bouret, P. *Adaptive test of homogeneity for a Poisson process.* Ann. Inst. H. Poincaré Probab. Statist., **47**(1), 176–213 (2011).

[18] Fromont, M., Laurent, B., Reynaud-Bouret, P. *The two-sample problem for Poisson processes: adaptive tests with a non-asymptotic wild bootstrap approach*, to appear in Annals of Statistics (2013).

[19] Gaïffas, S., Guilloux, A. *High-dimensional additive hazard models and the Lasso.* Electronic Journal of Statistics, **6**, 522–546 (2012).

[20] Giné, E., Latala, R., Zinn, J. *Exponential and Moment Inequalities for U-statistics.* High Dimensional Probability II - Progress in Probability, Birkhaüser, 13–38 (2000).

[21] Gusto, G., Schbath, S. *FADO: a statistical method to detect favored or avoided distances between motif occurrences using the Hawkes' model.* Statistical Applications in Genetics and Molecular Biology, **4**(1), Article 24 (2005).

[22] Hansen, N.R, Reynaud-Bouret, P., Rivoirard, V. *Lasso and probabilistic inequalities for multivariate point processes* Arxiv (2012).

[23] Hawkes, A. G., Oakes, D. *A cluster process representation of a self-exciting process.* J. Appl. Prob. **11**(3), 493–503 (1974).

[24] Houdré, C., Privault, N. *Concentration and deviation inequalities in infinite dimensions via covariance representations.* Bernoulli **8**(6), 697–720 (2002).

[25] Houdré, C., Reynaud-Bouret, P. *Exponential inequalities, with constants, for U-statistics of order two.* Stochastic inequalities and applications, Progr. Probab., 56 Birkhäuser, Basel, 55–69 (2003).

[26] Houdré, C., Marchal, P., Reynaud-Bouret, P. *Concentration for norms of infinitely divisible vectors with independent components.* Bernoulli, **14**(4), 926–948 (2008).

[27] Ingster, Yu.I., and Kutoyants, Yu.A. *Nonparametric hypothesis testing for intensity of the Poisson process.* Math. Methods Statist., **16** (3), 217–245 (2007).

[28] Kingman, J.F.C. *Poisson processes.* Oxford studies in Probability (1993).

[29] Klein, T. ,Rio, E. *Concentration around the mean for maxima of empirical processes.* Ann. Proba., **33**(3), 1060–1077 (2005).

[30] Ledoux, M. *On Talagrand inequalities for product mesures.* ESAIM PS, **1**, 95-144 (1996).

[31] Mallows, C.L. *Some comments on C_p.* Technometrics, **15**, 661–675 (1973).
[32] Massart, P. *Concentration inequalities and model selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Springer, Berlin (2007).
[33] Massart, P. *About the constants in Talagrand's concentration inequalities for empirical processes.* Ann. Proba., **28**(2), 863–884 (2000).
[34] Merton, R.C. *Option pricing when underlying stock returns are discontinuous.* Working paper Sloan School of Management, 787–795 (1975).
[35] Ogata, Y., *Statistical models for earthquakes occurrences and residual analysis for point processes.* Journal of the American Statistical Association, **83**(401), 9–27 (1988).
[36] Ogata, Y., Akaike, H. *On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes.* Journal of the Royal Statistical Society, Series B, **44**(1),102–107 (1982).
[37] Ozaki, T. *Maximum likelihood estimation of Hawkes' self-exciting point processes.* Ann. Inst. Statist. Math., **31**(B), 145–155 (1979).
[38] Papangelou, F.,*Integrability of expected increments of point processes and a related random change of scale*, Trans. Amer. Math. Soc., **165**,483–506 (1972).
[39] Pernice, V., Staude, B., Cardanobile, S., Rotter, S., *How structure determines correlations in neuronal networks.*, PLoS Computational Biology, 85:031916 (2012).
[40] Reinert, G., Schbath, S., Waterman, M.S. *Probabilistic and Statistical Properties of Words: An Overview.* Journal of Computational Biology, **7**(1–2), 1–46 (2000).
[41] Reynaud-Bouret, P. *Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities.* Probab. Theory Related Fields, **126** (1), 103–153 (2003).
[42] Reynaud-Bouret, P. *Compensator and exponential inequalities for some suprema of counting processes.* Statistics and Probability Letters, **76**(14), 1514–1521 (2006).
[43] Reynaud-Bouret, P. *Penalized projection estimators of the Aalen multiplicative intensity.* Bernoulli, **12**(4), 633–661 (2006).
[44] Reynaud-Bouret, P., Roy, E. *Some non asymptotic tail estimates for Hawkes processes.* Bulletin of the Belgian Mathematical Society-Simon Stevin, **13**(5), 883–896 (2007).
[45] Reynaud-Bouret, P., Rivoirard, V. *Near optimal thresholding estimation of a Poisson intensity on the real line.* Electronic Journal of Statistics, **4**, 172–238 (2010).
[46] Reynaud-Bouret, P., Schbath, S. *Adaptive estimation for Hawkes processes; application to genome analysis.* Ann. Statist., **38**(5), 2781–2822 (2010).
[47] Reynaud-Bouret, P., Rivoirard, V., Tuleau-Malot, C. *Adaptive density estimation: a curse of support?* J. Statist. Plann. Inference, **141**, 115–139 (2011).
[48] Reynaud-Bouret, P., Tuleau-Malot, C. , Rivoirard, V. and Grammont, F. *Spike trains as (in)homogeneous Poisson processes or Hawkes processes: non parametric estimation and goodness-of-fit tests.* Hal (2013).
[49] Samson, P.-M. *Concentration of measure inequalities for Markov chains and ϕ-mixing processes.* Ann. Probab., **28**(1), 416–461 (2000).
[50] Sansonnet, L. *Wavelet thresholding estimation in a Poissonian interactions model with application to genomic data.* to appear in Scandinavian Journal of Statistics (2012).
[51] Sansonnet, L., Tuleau-Malot, C. *A model of Poissonian interactions and detection of dependence.* Arxiv (2013).
[52] Shorack, G.R., Wellner, J.A. *Empirical processes with applications to statistics.*Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York (1986).
[53] Talagrand, M. *New concentration inequalities in product spaces.* Invent. Math., **126**(3), 505–563 (1996).
[54] Uhler, R.S. and Bradley, P. G. *A Stochastic Model for Determining the Economic Prospects of Petroleum Exploration Over Large Regions.* Journal of the American Statistical Association, **65**(330), 623–630 (1970).
[55] van de Geer, S. *Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes.* The Annals of Statistics, **23**(5), 1779–1801 (1995).
[56] Vere-Jones, D., Ozaki, T. *Some examples of statistical estimation applied to earthquake data.* Ann. Inst. Statist. Math., **34**(B), 189–207 (1982).
[57] Watson, G.S. *Estimating the intensity of a Poisson process.* Applied time series analysis, 1st proceeding, Tulsa, 1976, 325–345 (1978).
[58] Willett, R.M. Nowak, R.D. *Multiscale Poisson Intensity and Density Estimation.* IEEE Transactions on Information Theory, **53**(9), 3171–3187 (2007).
[59] Wu, L. *A new modified logarithmic Sobolev inequality for Poisson point process and several applications.* Probab. Theory Related Fields, **118**(3), 427–438 (2000).